

Discovery of missing disease spreader

Yoshiharu Maeno

Social Design Group

email: maeno.yoshiharu@socialdesigngroup.com

October 14, 2011

Abstract

Is it possible to discover a local outbreak of an infectious disease in a region for which data are missing, but which is at work as a disease spreader? Node discovery for the spread of an infectious disease is defined as discriminating between the nodes which are neighboring to a missing disease spreader node, and the rest, given a dataset on the number of cases. The spread is described by stochastic differential equations. A perturbation theory quantifies the impact of the missing spreader on the moments of the number of cases. Statistical discriminators examine the mid-body or tail-ends of the probability density function, and search for the disturbance from the missing spreader. They are tested with computationally synthesized datasets, and applied to the SARS outbreak and flu pandemic.

1 Introduction

No sooner had a new year begun in 2003 than citizens were seized with panic in Guangdong in south China. Hundreds were suffered from a pneumonia-like strange disease, some of which had been dead. Both Chinese government and Chinese media remained silent all the time as to the risk of a possible epidemic. No one in the rest of the world knew there was any real cause for alarm. But in March, local outbreaks of a mysterious disease were reported in Hong Kong and Southeast Asian countries. The World Health Organization (WHO) issued a global alert. Even then, health authorities could not reveal where the disease had come from. This story at the onset of the Severe Acute Respiratory Syndrome (SARS) outbreak poses an interesting question. Is it possible to discover the presence of a missing disease spreader from the surveillance records on the cases in other regions? This study addresses such a node discovery problem for the spread of an infectious disease.

The spread of an infectious disease is a stochastic phenomenon in a spatially heterogeneous medium. Many mathematical models of disease transmission [Riley 2007] rely on an epidemiological compartment model and a meta-population network model in formulating stochasticity and spatial heterogene-

ity [Dangerfield 2009], [Simões 2008]. These models are described by a set of stochastic differential equations [Hufnagel 2004]. The analysis of the spread includes many tasks from reproduction to prediction. Reproduction of the actual spread [Christensen 2010], [Isham 2010] is essential in understanding the role of a so-called super-spreader [Fujie 2007], [Small 2006] and epidemiological thresholds [Parshani 2010], [Colizza 2007]. This is a forward analysis. An inverse analysis includes estimating the transmission parameters from observation [Walker 2010], [Keeling 2004]. Network profiling estimates the effectively decisive topology of a transportation network which governs the spread [Maeno 2010]. Network inference in a communication network [Rabbat 2008] is a similar problem. Statistical learning and computational Monte-Carlo simulation contribute to developing a reliable bio-surveillance system in a noisy environment [Reis 2003], detecting abnormal events [Takeuchi 2006] as an omen of the outbreak [Lu 2010], and predicting the spread to aid the health authorities in containing the outbreak [Colizza 2006]. None of these, however, addresses the problem. Node discovery problem for a social network [Maeno 2009] is discovering a person who does not appear in the logs on communication, but actually influential to others in an organization. The substantial nature of the problem is the same as this. But the mathematical solution should be different entirely because of the difference in the mechanism of the spread.

In this study, a perturbation theory quantifies how the perturbation which a missing spreader node exerts disturbs the growth of the number of cases. The missing spreader may be the place where the first patient appears (an index node) or not (an intermediate node). Its presence impedes the reproduction of the observed spread by an unperturbed mathematical model. The irreproducibility is rather the clue to solve the problem. Two statistical discriminators are invented. Their role is discriminating between the nodes which are neighboring to the missing spreader node, and the rest (non-neighboring nodes), given a dataset. It is a collection of the time sequence data on the number of cases, or on the cumulative number of new cases, at the nodes within the scope of surveillance in the early growth phase of the spread. The network topology and transmission parameters may be given as an input to the discriminators. Or they may be unknown and must be estimated. The discriminators examine the mid-body or the tail-ends of the probability density function of the number of cases, and search for disturbed time sequence data to which the perturbation gives rise. The mid-body discriminator is founded on the Kolmogorov-Smirnov test. The tail-end discriminator is founded on the Chauvenet rejection test. The discriminators are tested with a number of computationally synthesized datasets, and applied to the WHO datasets on the SARS outbreak and on the flu pandemic (H1N1 swine influenza A) in 2009.

2 Problem

The mathematical model of the spread in this study is a special case of a stochastic reaction-diffusion process, the integration of a standard epidemio-

logical SIR compartment model and a meta-population network model. The meta-population network model [Baronchelli 2008] sub-divides the entire population into distinct sub-populations in many geographical regions. The geographical regions are the nodes n_i ($i = 0, 1, \dots$). The transportation between two regions is a pair of unidirectional links. The adjacency matrix \mathbf{l} , whose i -th row and j -th column element is l_{ij} , determines the network topology. If a link from n_i to n_j is present, l_{ij} is 1. If absent, l_{ij} is 0. The parameter γ is a matrix whose i -th row and j -th column element γ_{ij} is the probability at which a person moves from n_i to n_j per a unit time. In many application areas, an empirical law $\gamma_{ij} = \Gamma_{ij}(\mathbf{l})$ determines γ as a function of \mathbf{l} . In the following, γ and \mathbf{l} are interchangeable.

In the SIR compartment model [Keeling 2008] the state of a person changes from a susceptible state, through an infectious state, to a *removed (recovered)* state. The quantity $S_i(t)$ is the number of susceptible persons at a node n_i at time t . $I_i(t)$ is the number of infectious persons. $R_i(t)$ is the number of recovered persons. $J_i(t)$ is the cumulative number of new cases until t . The parameter α represents the probability at which an infectious person contacts a person and infect the person per a unit time. The parameter β represents the probability at which an infectious person recovers per a unit time. These transmission parameters do not depend on time and sub-populations. The reproductive ratio r is defined by $r = \alpha/\beta$ [Lipsitch 2003]. Movement, infection and recovery are Markovian stochastic processes which are governed by $\boldsymbol{\theta} = \{\alpha, \beta, \mathbf{l}\}$.

Node discovery is a problem to determine whether the nodes n_0, n_1, \dots, n_{N-1} which appear in a given dataset is neighboring to a missing spreader node n_N or not. The neighboring nodes are those having links to n_N . The dataset is either $I_i(t_d)$, or $\Delta J_i(t_d) = J_i(t_{d+1}) - J_i(t_d)$, where $i = 0, 1, \dots, N-1$ denotes the nodes, and $d = 0, 1, \dots, D-1$ denotes observations. The interval between observations is $\Delta t = t_{d+1} - t_d$. The number of data is ND . The network topology and the transmission parameters $\boldsymbol{\theta} = \{\alpha, \beta, l_{ij}\}$ for $i, j = 0, 1, \dots, N-1$ are given under some conditions, and unknown under the other conditions. No other information is available. The node n_N is either present or absent. If present, it is either an index node or an intermediate node. The parameters l_{iN} and l_{Nj} are unknown under any conditions. An empirical law $\gamma_{ij} = \Gamma_{ij}(\mathbf{l})$ is postulated in this study. Without the law, when $\boldsymbol{\theta}$ were unknown, the continuous parameter γ would be estimated, rather than the binary parameter \mathbf{l} .

3 Method

3.1 Probability density function

The time evolution of $I_i(t)$, or $J_i(t)$ is given by Langevin equations [Hufnagel 2004]. They are a system of stochastic differential equations [Kampen 2007]. The ensemble of an infinite number of sample trajectories $I_i(t)$ is equivalent to the time dependent joint probability density function $P(\mathbf{I}, t)$ of the probability variables $\mathbf{I} = (I_0, I_1, \dots)$. Stochastic differential equations for $I_i(t)$ are

converted to a partial differential equation for $P(\mathbf{I}, t)$. This Fokker-Planck equation [Kampen 2007] is converted to ordinary differential equations to calculate the moments of I_i one order after another. The first order moments $m_i(t|\boldsymbol{\theta}) = \int I_i P(\mathbf{I}, t) d\mathbf{I}$ are the mean of I_i at t . The second order moments $v_{ij}(t|\boldsymbol{\theta})$ are the covariance about the mean between I_i and I_j . The third order moments $s_{ijk}(t|\boldsymbol{\theta})$ are the skewness about the mean among I_i , I_j , and I_k . The fourth order moments $\kappa_{ijkl}(t|\boldsymbol{\theta})$ are the kurtosis about the mean among I_i , I_j , I_k , and I_l . The fifth and higher order moments are not analyzed. The formulae for the moments are listed in Appendix A. The observations $I_i(t_d)$ at t_d are the initial condition to obtain the moments at t_{d+1} . In the following, the lower order moments refer to the mean and variance, and the higher order moments refer to the skewness and kurtosis. For small Δt , $s_{ijk}(t_{d+1}|\boldsymbol{\theta}), \kappa_{ijkl}(t_{d+1}|\boldsymbol{\theta}) \ll m_i(t_{d+1}|\boldsymbol{\theta}), v_{ij}(t_{d+1}|\boldsymbol{\theta})$. The probability density function is a multi-variate normal distribution in eq.(1). The i -th element of the row vector $\mathbf{m}(t_{d+1}|\boldsymbol{\theta})$ is $m_i(t_{d+1}|\boldsymbol{\theta})$. The i -th row and j -th column element of the matrix $\mathbf{v}(t_{d+1}|\boldsymbol{\theta})$ is $v_{ij}(t_{d+1}|\boldsymbol{\theta})$.

$$P(\mathbf{I}, t_{d+1}) \approx P_N(\mathbf{I}; \mathbf{m}(t_{d+1}|\boldsymbol{\theta}), \mathbf{v}(t_{d+1}|\boldsymbol{\theta})). \quad (1)$$

$P(\mathbf{I}, t)$ depends on time and many probability variables. Analyzing a dataset is an involved task. Time dependent conditional z -score for I_i resolves the involvedness. The variables I_i at t_{d+1} are converted to the variables z_i defined by eq.(2).

$$z_i = \frac{I_i - m_i^C(t_{d+1}|\boldsymbol{\theta})}{\sqrt{v_{ii}^C(t_{d+1}|\boldsymbol{\theta})}}. \quad (2)$$

In eq.(2), the mean $m_i^C(t_{d+1}|\boldsymbol{\theta})$ and the variance $v_{ii}^C(t_{d+1}|\boldsymbol{\theta})$ are conditioned on the observation for the rest of the variables $\mathbf{I}_{\bar{i}} = (I_0, \dots, I_{i-1}, I_{i+1}, \dots, I_{N-1})$ at t_{d+1} . Generally, given $\mathbf{I}_{\bar{i}}$, the conditional probability density function for I_i is a uni-variate normal distribution $P_N(I_i; m_i^C, v_{ii}^C)$ if \mathbf{I} obeys a multi-variate normal distribution $P_N(\mathbf{I}; \mathbf{m}, \mathbf{v})$. The conditional mean $m_i^C(t_{d+1}|\boldsymbol{\theta})$ is given by eq.(3). It is a sum of the unconditional mean m_i and a term dependent on the difference between the observation and the expected value $(\mathbf{I} - \mathbf{m})$ for the rest of the variables at t_{d+1} . The column vector $(\mathbf{I} - \mathbf{m})^T$ is a transpose of a row vector $\mathbf{I} - \mathbf{m}$.

$$m_i^C(t_{d+1}|\boldsymbol{\theta}) = m_i(t_{d+1}|\boldsymbol{\theta}) + \mathbf{v}_{i\bar{i}}(t_{d+1}|\boldsymbol{\theta}) \mathbf{v}_{\bar{i}\bar{i}}(t_{d+1}|\boldsymbol{\theta})^{-1} (\mathbf{I}_{\bar{i}} - \mathbf{m}_{\bar{i}}(t_{d+1}|\boldsymbol{\theta}))^T. \quad (3)$$

The conditional variance $v_{ii}^C(t_{d+1}|\boldsymbol{\theta})$ is a Shur compliment of $\mathbf{v}_{\bar{i}\bar{i}}$ in \mathbf{v} . It is given by eq.(4). Because the observation reduces uncertainty, v_{ii}^C is smaller than v_{ii} by the amount determined by the second term.

$$v_{ii}^C(t_{d+1}|\boldsymbol{\theta}) = v_{ii}(t_{d+1}|\boldsymbol{\theta}) - \mathbf{v}_{i\bar{i}}(t_{d+1}|\boldsymbol{\theta}) \mathbf{v}_{\bar{i}\bar{i}}(t_{d+1}|\boldsymbol{\theta})^{-1} \mathbf{v}_{\bar{i}i}(t_{d+1}|\boldsymbol{\theta}). \quad (4)$$

In eq.(3) and (4), the unconditional mean is partitioned into the i -th element and a row vector. The unconditional covariance is partitioned into four sub-matrices (1×1 , $1 \times (N-1)$, $(N-1) \times 1$, $(N-1) \times (N-1)$ matrices). These

are given by eq.(5).

$$\mathbf{m} = (m_i, \mathbf{m}_{\bar{i}}), \mathbf{v} = \begin{pmatrix} v_{ii} & \mathbf{v}_{i\bar{i}} \\ \mathbf{v}_{\bar{i}i} & \mathbf{v}_{\bar{i}\bar{i}} \end{pmatrix} \quad (5)$$

The non-uniform growth at different nodes at different times is absent in z_i . Eq.(1) becomes eq.(6), which is valid for the all nodes n_i at any time t as far as the approximation that \mathbf{I} obeys a multi-variate normal distribution holds true.

$$P(z_i, t_{d+1} | \mathbf{I}_{\bar{i}}, \boldsymbol{\theta}) = P_N(z_i; 0, 1). \quad (6)$$

In the above discussion, it is assumed that $\boldsymbol{\theta}$ is known and $I_i(t_d)$ is given as a dataset. If $\boldsymbol{\theta}$ is unknown, the network topology and transmission parameters must be estimated from a given dataset. A well-known statistical technique for this purpose is the maximal likelihood estimation or the maximal a posteriori probability estimation [Hastie 2001]. The true parameter $\boldsymbol{\theta}$ is substituted by the estimator $\hat{\boldsymbol{\theta}}$. If $\Delta J_i(t_d)$ is given, instead of $I_i(t_d)$, $\Delta J_i(t_d)$ is converted to $I_i(t_d)$ by eq.(7). The all formulae for $I_i(t_d)$ are applicable then.

$$I_i(t_d) \approx \frac{\Delta J_i(t_d)}{\alpha \Delta t}. \quad (7)$$

Eq.(7) is a discrete-time approximation to the stochastic differential equation. The fluctuation terms are simply discarded. The motivation to use eq.(7) is as follows. The model in this study is a hidden Markovian model in the sense that the observation $\Delta J_i(t_d)$ is determined by the unobserved states of variables $I_i(t)$ in a Markovian stochastic process. It is known generally intractable to estimate the model with many continuous-time dependent (so-called heteroskedastic) continuous latent variables. Such an approximation as eq.(7) is critical in the estimation. If the true value α in eq.(7) is unknown, it is substituted by the estimator $\hat{\alpha}$. The mathematical details for these estimation and conversion are presented in [Maeno 2010].

3.2 Perturbation theory

A perturbation theory quantifies the impact of a missing spreader node on its neighboring nodes and non-neighboring nodes. Let us investigate a simple network which consists of a missing spreader node n_s , a neighboring node n_n which is connected to n_s , and a non-neighboring node n_a which stays apart from n_s . That is, $N = 2$ and $n_2 = n_s$. The three nodes are connected with two links between n_n and n_a , and n_n and n_s . Assume that $\gamma_{na} = \gamma_{an} = \gamma$, $\gamma_{ns} = \gamma_{sn} = \gamma'$, and $\gamma_{as} = \gamma_{sa} = 0$. The presence of n_s means $\gamma' > 0$. Given $\boldsymbol{\theta} = \{\alpha, \beta, \gamma\}$, how does non-zero γ' disturb the moments of I_n and I_a ?

The formulae for the moments in this network are listed in Appendix B. The mean for n_n changes in the direction determined by the sign of $I_n(t_d) - I_s(t_d)$. The variance increases in proportion to γ' . The terms dependent on $O(\gamma'^2)$ appear in the formulae for the skewness and kurtosis. The mean and variance

of the z -score for n_n are given by eq.(8) and (9). The quantity $\langle z \rangle$ is the average over the observations at different times. In eq.(2), the disturbance is included in the observation I_i whose moments are $m_i(t_{d+1}|\boldsymbol{\theta}, \gamma')$ and $v_{ii}(t_{d+1}|\boldsymbol{\theta}, \gamma')$. On the other hand, nothing can be assumed about the presence of n_s in calculating m_i^C and v_{ii}^C . These are the values predicted from $m_i(t_{d+1}|\boldsymbol{\theta}, 0)$ and $v_{ii}(t_{d+1}|\boldsymbol{\theta}, 0)$ when $\gamma' = 0$, given $\boldsymbol{\theta} = \{\alpha, \beta, \gamma\}$.

$$\langle z_n \rangle = -\frac{\gamma'(I_n - I_s)}{\sqrt{(\alpha + \beta)I_n + \gamma(I_n + I_a)}}\sqrt{\Delta t}. \quad (8)$$

$$\langle (z_n - \langle z_n \rangle)^2 \rangle = 1 + \frac{\gamma'(I_n + I_s)}{(\alpha + \beta)I_n + \gamma(I_n + I_a)}. \quad (9)$$

As γ' increases, the difference between the probability density function of z_n and the standardized normal distribution becomes more significant. In contrast, the mean, variance, and skewness for n_a do not change at all. Interestingly, the kurtosis increases in proportion to γ' . The presence of n_s disturbs the fourth and higher order moments for n_a . In terms of the z -score for n_a , $\langle z_a \rangle = 0$ and $\langle (z_a - \langle z_a \rangle)^2 \rangle = 1$. The coupling with a spreader and non-neighboring nodes emerges at this order. But such a signal from non-zero γ' may be too weak to detect from n_a . The above discussion implies that the normality of the probability density function for neighboring nodes is vulnerable to the perturbation which a missing spreader node exerts, but that the impact on the normality for non-neighboring nodes is not salient. This is the basis to discriminate between the neighboring nodes and non-neighboring nodes statistically.

Let us extend the theory to multiple non-neighboring nodes. The entire set of nodes is treated as a big node n_a . The number of infectious persons at the individual nodes is roughly the same as I_n . Assume there are k such nodes. $I_a \approx kI_n$ holds. If n_s is an index node, $I_s \gg I_n$ holds. The variance of z_n in eq.(9) is sufficiently large as a signal when $k < \gamma'I_s/\gamma I_n$. Any k satisfy this criterion unless γ' is extremely small. If n_s is an intermediate node, $I_s \approx I_n$ holds. The variance is large enough when $k < (2\gamma' - \gamma - \alpha - \beta)/\gamma$. Small k , rather than small N , seems a prerequisite for discrimination.

When a household or a hospital ward form a sub-population, the population of a node is very small. Most of the sub-population are likely to get infected immediately at the onset of infection. The day-by-day fluctuating growth of the number of patients, in which the trace of perturbation is left, will not be observed. The formulae in this study may not work under this condition. Moreover as an extreme, a social network model, where an individual is a node, is suitable to analyze the transmission of such diseases as Acquired Immune Deficiency Syndrome (AIDS) [Potterat 1999]. Formulating a stochastic model and perturbation theory for the social network model is beyond the scope of this study.

3.3 Statistical discriminator

3.3.1 Mid-body discriminator

The mid-body discriminator is founded on the Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test [Press 2007] estimates the minimal distance between two cumulative density functions. In many applications, the test is used for a one sample hypothesis testing where the null hypothesis is that the cumulative density function drawn from a dataset empirically is the same as a given reference cumulative density function. A typical reference function is a cumulative normal distribution with a given mean and variance. The test is more sensitive to the difference in the mid-bodies of the cumulative density functions than that in the tail-ends. In other words, it is more sensitive to the difference in the lower order moments. Let $z(d)$ denote multiple observations ($d = 0, 1, \dots, D-1$) for a single probability variable z . The empirical cumulative density function $F(z)$ is given by eq.(10). The function $H(x)$ for a real argument x is a Heaviside's function whose value is 1 when $x > 0$, and 0 when $x < 0$.

$$F(z) = \frac{1}{D} \sum_{d=0}^{D-1} H(z - z(d)). \quad (10)$$

The test statistic T is given by eq.(11). $F_R(z)$ is the reference cumulative density function which z obeys in the hypothesis. This is the minimal distance between the cumulative density functions. The supremum $\sup x$ for a real variable x is the least upper bound of x .

$$T = \sqrt{D-1} \sup_z |F(z) - F_R(z)|. \quad (11)$$

The null hypothesis is rejected at the significance level of a when $T > K_a$ where $K_a = K^{-1}(1-a)$. $K(x)$ is a cumulative density function of the Kolmogorov distribution for a probability variable $x > 0$. $K_a = 1.38$ for $a = 5\%$, and $K_a = 1.63$ for $a = 1\%$. Eq.(10) is applicable to calculating the empirical cumulative density function $F_i(z)$ of the z -score for n_i in eq.(2) from the observation of I_i at $t = t_d$ ($d = 0, 1, \dots, D-1$). Because of the property in eq.(6), the reference cumulative density function is the cumulative standard normal distribution $F_R(z) = (1 + \text{erf}(z/\sqrt{2}))/2$. The test statistic T in eq.(11) for n_i becomes T_i in eq.(12).

$$T_i = \sqrt{D-1} \sup_z \left| \frac{1}{D-1} \sum_{d=0}^{D-2} H(z - z_i(t_{d+1})) - \frac{1 + \text{erf}(z/\sqrt{2})}{2} \right|. \quad (12)$$

It is not obvious which value of the threshold K_a is the most suitable because the normality for non-neighboring nodes and non-normality for neighboring nodes are just an approximation. On the other hand, it is difficult to derive an analytical formula for an appropriate discrimination threshold as a function of the parameters and experimental conditions. Searching for the threshold

T^* experimentally is rather practical. If $T_i > T^*$, the mid-body discriminator determines that the perturbation from an unknown origin disturbs I_i , and consequently that n_i is neighboring to a missing spreader node.

3.3.2 Tail-end discriminator

The tail-end discriminator is founded on the Chauvenet rejection test. The Chauvenet rejection test [Taylor 1996] detects an outlier in a given dataset. It was invented as a criterion to assess statistically whether particular one-dimensional numerical data are likely to be spurious or not, and is used widely in experimental physics and chemistry today. First, the mean and variance of a given dataset are calculated. The probability at which the individual datum is obtained under the calculated mean and variance is evaluated. Then, the datum is considered to be an outlier if the product of the probability and the number of data in the dataset is less than a given threshold. The threshold is 0.5 in the conventional Chauvenet rejection test. For example, the data are spurious if the probability is less than 0.05 when the number of data is 10.

The test statistic L is the likelihood function [Hastie 2001], which is the conditional probability of the data as a function of the parameters. The conditional probability will be noticeably small if the data are spurious, that is, the data lies in the tail-ends. The test is sensitive to the anomaly in the higher order moments. The test statistic L_i of the z -score for n_i is defined by eq.(13).

$$L_i = \frac{1}{D-1} \sum_{d=0}^{D-2} \ln P(z_i, t_{d+1} | \mathbf{I}_i, \boldsymbol{\theta}) = \frac{1}{D-1} \sum_{d=0}^{D-2} \ln P_N(z_i; 0, 1). \quad (13)$$

According to the conventional Chauvenet rejection test, the discrimination threshold for the test statistic is $C = \ln 0.05 = -3$ when $N = 10$. Recall the discussion on the significance of the absolute value of K_a for the mid-body discriminator. Instead of relying on C in the conventional Chauvenet rejection test, searching for an appropriate discrimination threshold L^* experimentally is rather appropriate. If $L_i < L^*$, the tail-end discriminator determines that a node n_i is a neighboring node.

4 Experiment

4.1 Computationally synthesized dataset

The performance of the discriminators is studied with a number of computationally synthesized test datasets. The test datasets are synthesized by numerical integration [Kloeden 1992] of a set of Langevin equations in eq.(17) for random network topologies and transmission parameters. The network is a Erdős-Rényi model. The probability at which a link is present between n_i and n_j ($l_{ij} = l_{ji} = 1$) is a constant $\langle k_i \rangle / (N - 1)$. The nodal degree of a node n_i is given by $k_i = \sum_j l_{ij}$. The average over the all nodes is $\langle k_i \rangle$.

Table 1: Four possible outcomes from a discriminator.

| | | Actual value | |
|----------------------|-----------------|--------------------------------|--------------------------------|
| | | Neighboring | Non-neighboring |
| Discriminator output | Neighboring | True positive N_{TP} | False negative N_{FP} |
| | Non-neighboring | False positive N_{FN} | True negative N_{TN} |

It is postulated that the total number of persons who moves from n_i to n_j per a unit time is proportional to $\sqrt{k_i k_j}$ if a link is present. This is valid generally for the world-wide airline transportation network [Barrat 2004]. Eq.(14) determines γ_{ij} as a function of \mathbf{l} . The fraction of persons who outgoes per a unit time is a constant γ over the network.

$$\gamma_{ij} = \Gamma_{ij}(\mathbf{l}) = \frac{l_{ij} \sqrt{k_i k_j}}{\sum_j l_{ij} \sqrt{k_i k_j}} \gamma. \quad (14)$$

It is also postulated that the initial population $P_i(0) = S_i(0) + I_i(0) + R_i(0)$ of a node n_i is proportional to the total number of persons who outgoes from the node per a unit time [Maeno 2010]. $P_i(0)$ is given by eq.(15). The total population is $P = 10^6 N$ in the experiment. This is relevant in synthesizing the test datasets, but not necessary in discrimination.

$$P_i(0) = \frac{\sum_j l_{ij} \sqrt{k_i k_j}}{\sum_{i,j} l_{ij} \sqrt{k_i k_j}} P. \quad (15)$$

A receiver operating characteristic curve [Fawcett 2006] is drawn to evaluate the accuracy of discrimination. In signal processing, a receiver operating characteristic curve is a plot of the true positive ratio R_{TP} on the vertical axis and the false positive ratio R_{FP} on the horizontal axis for a binary discriminator as its discrimination threshold is varied. It is generally a concave function. The discrimination threshold is T^* for the mid-body discriminator, and L^* for the tail-end discriminator. There are four possible outcomes from the discriminator. They are summarized in Table 1. True positive and true negative are right answers, but false positive and false negative are wrong answers. The number of the true positive is N_{TP} . The number of nodes is $N = N_{\text{TP}} + N_{\text{FN}} + N_{\text{FP}} + N_{\text{TN}}$. The ratios are defined by eq.(16). R_{TP} is called recall alternatively, and R_{FP} is called fallout.

$$R_{\text{TP}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}, \quad R_{\text{FP}} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}}. \quad (16)$$

The ratios take the value between 0 and 1. If the discriminator works ideally excellently, $N_{\text{TP}} = k_N$, $N_{\text{FN}} = 0$, $N_{\text{FP}} = 0$, and $N_{\text{TN}} = N - k_N$. The curve for the ideal discriminator degenerates to the upper left corner $(R_{\text{FP}}, R_{\text{TP}}) = (0, 1)$.

The curve for a random discriminator is a straight line $R_{\text{TP}} = R_{\text{FP}}$ between $(0,0)$ and $(1,1)$. The curve for a more excellent discriminator comes closer to the upper left corner. The closeness of $R_{\text{TP}} - R_{\text{FP}}$ to its ideal value of 1 is a scalar indicator of accuracy. It is used as an objective function to search for the optimal thresholds L^* and T^* experimentally.

Figure 1 **a, b** shows the receiver operating characteristic curves for $N = 10$ when the missing spreader node n_{10} is an index node. The tail-end discriminator works excellently both when θ is given and unknown. The best performance is $R_{\text{TP}} - R_{\text{FP}} \approx 1$. The mid-body discriminator is the most suitable when θ is given. The best performance is $R_{\text{TP}} - R_{\text{FP}} \approx 1$ for given θ , and ≈ 0.2 for unknown θ . Figure 1 **c, d** shows the curves for $N = 30$ when the missing spreader node n_{30} is an index node. Figure 1 **e, f** shows the curves for $N = 10$ when the missing spreader node n_{10} is an intermediate node. Discrimination is not as excellent as that in **a, b**, but excellent moderately. The best performance of the tail-end discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.9$ for given θ , and ≈ 0.7 for unknown θ . The best performance of the mid-body discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.65$ for given θ . The intermediate node and its neighboring nodes look alike as sinks of infected travelers indistinguishably. This is in strong contrast to the index node which is a salient source of them.

When θ is unknown, the estimation searches for a network topology of N nodes and transmission parameters, whose behavior bears the closest resemblance to I_0, I_1, \dots, I_{N-1} in the actual network topology of $N + 1$ nodes. In the simple network in 3.2, this results in $m_i(t_{d+1}|\hat{\theta}, 0) \sim m_i(t_{d+1}|\theta, \gamma')$ and $v_{ij}(t_{d+1}|\hat{\theta}, 0) \sim v_{ij}(t_{d+1}|\theta, \gamma')$, consequently $\langle z_n \rangle \sim 0$ and $\langle (z_n - \langle z_n \rangle)^2 \rangle \sim 1$, rather than eq.(8) and (9). This is confirmed by measuring the moments with the datasets. The measured moments are shown in Appendix C. When a missing spreader node is absent, the moments for given θ are almost the same as those for the estimator $\hat{\theta}$. The estimator reproduces the spread accurately. The moments are $\langle m_i \rangle \approx 0$ and $\langle v_{ii} \rangle = 1$. On the other hand, the difference in every moment between neighboring nodes and non-neighboring nodes for the estimator $\hat{\theta}$ is much smaller than that for given θ when the missing spreader node is an index node. The difference in $\langle m_i \rangle$, and that in $\langle v_{ii} \rangle$ are so small that n_n can not be distinguished from n_a by examining the lower order moments. This is the reason why the mid-body discriminator is not suitable if θ is unknown. Furthermore, the difference in every moment between neighboring nodes and non-neighboring nodes is particularly small when the missing spreader node is an intermediate node. Discovering a missing intermediate node is a hard problem.

Figure 2 **a, b** shows the curves when the number of data is $D = 33$. Accuracy is the same as that in Figure 1 **a, b**. It does not depend on these experimental conditions. Figure 2 **c, d** shows the curves when the reproductive ratio is $r = 8$, which is four times larger than the value in Figure 1 **a, b**. The best performance of the tail-end discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.95$ for given θ , and ≈ 0.85 for unknown θ . The best performance of the mid-body discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.95$ for given θ . There is a small degradation in accuracy,

but discrimination is still excellent. Figure 2 **e, f** shows the curves when the fraction is $\gamma = 0.4$, which is four times larger than the value in Figure 1 **a, b**. The accuracy is the same as that in Figure 1 **a, b**.

Figure 3 **a, b** shows the curves when $\Delta J_i(t_d)$, instead of $I_i(t_d)$, is given as a dataset. The missing spreader node n_{10} is an index node. The experimental conditions are the same as those in Figure 1 **a, b**. These two spreads are identical. Discrimination from $\Delta J_i(t_d)$ is not so excellent as than from $I_i(t_d)$ in Figure 1 **a, b**. The best performance of the tail-end discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.55$ for given θ , and ≈ 0.85 for unknown θ . The best performance of the mid-body discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.65$ for given θ . Figure 3 **c, d** shows the curves under the condition where the missing spreader node n_{10} is an intermediate node. The best performance of the tail-end discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.4$ both for given and unknown θ . The best performance of the mid-body discriminator is $R_{\text{TP}} - R_{\text{FP}} \approx 0.6$ for given θ . The performance of the mid-body discriminator does not change much between **a, b** and **c, d**.

Somehow these tendencies are contrary to the results so far. Recall that the fluctuation term in eq.(19) is discarded in the approximation of eq.(7). The term is more complex than a mere Gaussian white noise because its amplitude includes $\sqrt{I_i(t)}$, and the ensemble of sample trajectories $I_i(t)$ does not obey a multi-variate normal distribution in eq.(1) for large t . Eq.(7) is correct on the average, but the conversion of the higher order moments from $\Delta J_i(t_d)$ to $I_i(t_d)$ is inaccurately distortive. It is the reason why the tail-end discriminator is only moderately excellent, and the shape of the curve looks awkward when θ is given. On the other hand, when θ is unknown, the estimator $\hat{\theta}$ may be able to make up for the distortion in the higher order moments. The tail-end discriminator with the estimator $\hat{\theta}$ outperforms the discriminators for given θ . Even if the true parameters are given, adjusting the parameters is rather indispensable for excellent discrimination from $\Delta J_i(t_d)$. Substantial improvement of eq.(7) is beyond the scope of this study and for future challenge.

Figure 4 **a, b** shows $R_{\text{TP}} - R_{\text{FP}}$ as a function of the discrimination threshold L^* or T^* when θ is given. The rising edges of the three curves coincide with each other. $L^* = -3$ to -3.5 , and $T^* = 1.5$ to 2 are reasonable if there is no prior information on the presence or absence of a missing spreader node. If the assumption is grounded well that an index node is missing, there is a distinct optimal threshold $L^* = -4$, and $T^* = 2.1$. Figure 4 **c, d** shows $R_{\text{TP}} - R_{\text{FP}}$ when θ is unknown. Setting $L^* = -3$ to -3.5 is reasonable. The position of the falling edges of the curves in **c, d** are different from that in **a, b**. There is a distinct optimal threshold $L^* = -3.4$ for a missing index node.

Figure 5 **a, b** shows the optimal thresholds L^* and T^* as a function of the transmission parameters r (α and β) and γ . Figure 5 **c, d** shows the optimal thresholds L^* and T^* as a function of the number of nodes N . The optimal thresholds depend on r . This means that it is practical to find the optimal threshold experimentally, which is suitable for individual conditions. The dependence on γ and N is so small that the dependence can be ignored.

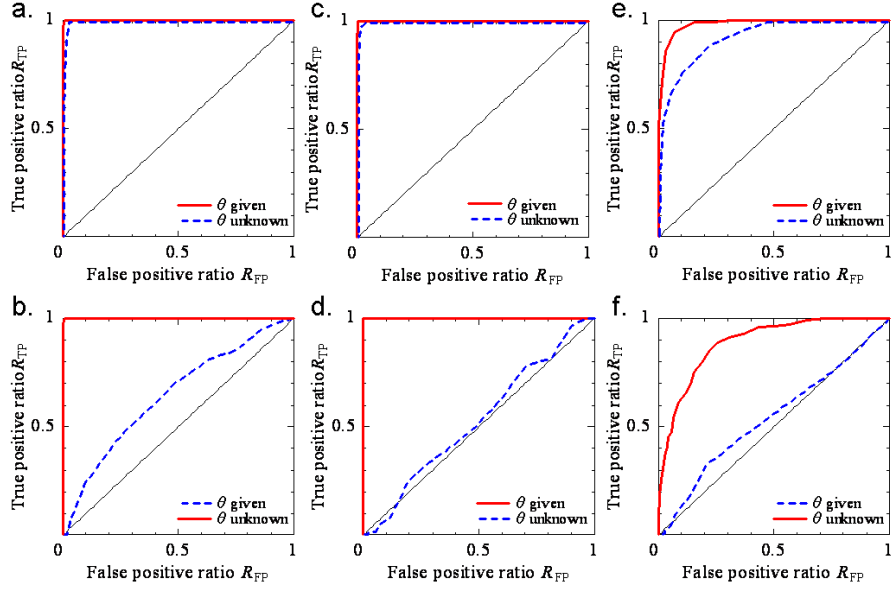


Figure 1: Receiver operating characteristic curves of the tail-end discriminator (a, c, e), and the mid-body discriminator (b, d, f). The parameters are $\langle k_i \rangle = 2$, $r = 2$ ($\alpha = 0.067$, $\beta = 0.033$), and $\gamma = 0.1$. $I_i(t_d)$ for $0 \leq d \leq 99$ ($D = 100$) with $\Delta t = 1$ is given as a dataset. a, b, $N = 10$, the missing spreader node n_{10} is an index node, and $I_{10}(0) = 200$. c, d, $N = 30$, n_{30} is an index node, and $I_{30}(0) = 200$. e, f, $N = 10$, n_{10} is not an index node but an intermediate node, and $I_0(0) = 200$. The curves are drawn from the trials for 100 different random network topologies.

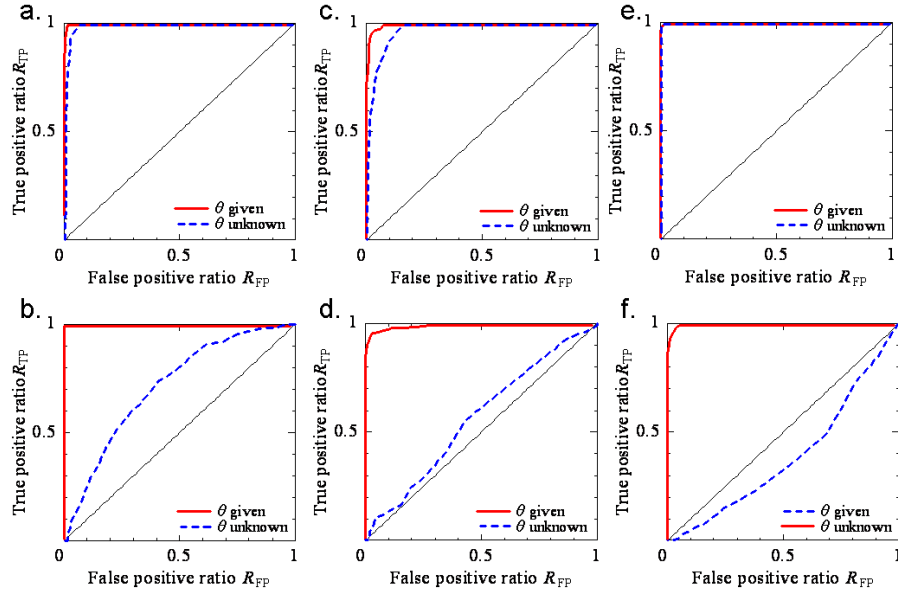


Figure 2: Receiver operating characteristic curves of the tail-end discriminator (a, c, e), and the mid-body discriminator (b, d, f). a, b, $I_i(t_d)$ for $0 \leq d \leq 32$ ($D = 33$) is given as a dataset. c, d, $r = 8$ ($\alpha = 0.089, \beta = 0.011$). e, f, $\gamma = 0.4$. The other experimental conditions are the same as those in Figure 1 a, b.

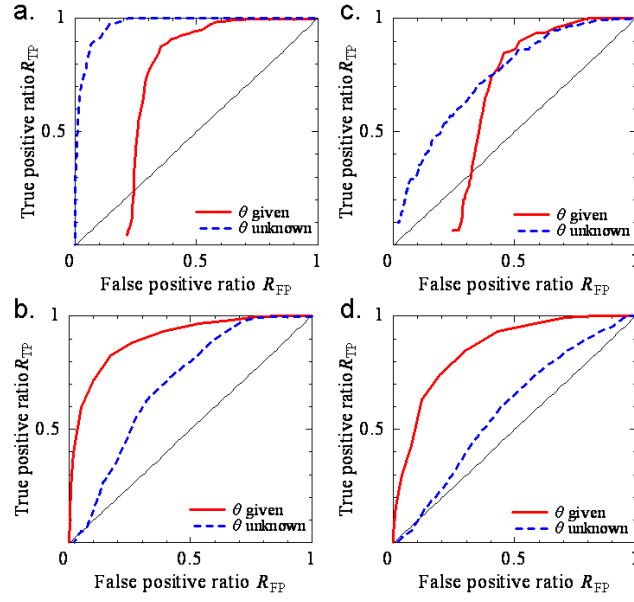


Figure 3: Receiver operating characteristic curves of the tail-end discriminator (a, c), and the mid-body discriminator (b, d). $\Delta J_i(t_d)$ for $0 \leq d \leq 99$ ($D = 100$) with $\Delta t = 1$ is given as a dataset. a, b, the missing spreader node n_{10} is an index node, and $I_{10}(0) = 200$. c, d, n_{10} is not an index node but an intermediate node, and $I_0(0) = 200$. The experimental conditions are the same as those in Figure 1 a, b, e, f.

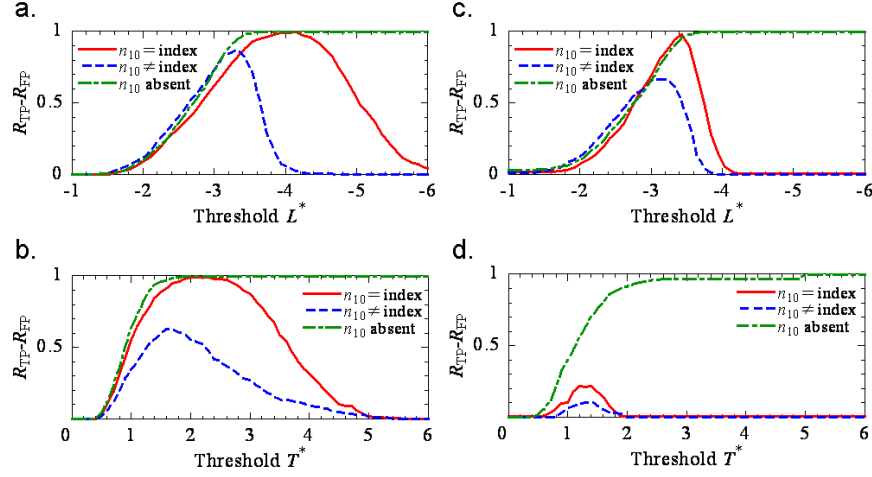


Figure 4: $R_{TP} - R_{FP}$ of the tail-end discriminator (a, c), and the mid-body discriminator (b, d) as a function of the discrimination thresholds. a, b, θ is given. c, d, θ is unknown and estimated from the dataset. The missing spreader node n_{10} is either an index node, an intermediate node, or absent. The experimental conditions are the same as those in Figure 1 a, b, e, f.

4.2 WHO dataset

4.2.1 SARS outbreak in 2003

SARS is a respiratory disease in humans caused by the SARS corona-virus. The epidemic of SARS appears to have started in Guangdong of south China in November 2002. SARS spread from the Guangdong to Hong Kong in early 2003 [Lipsitch 2003], and eventually nearly 40 countries around the world by July [Riley 2003]. WHO archives the cumulative number of reported probable cases of SARS¹. The dataset in the archive had been updated every day. It is a collection of time sequence data $J_i(t_d)$ with $\Delta t = 1$ day.

The target geographical regions in this study are those where five or more cases had been reported in a month since March 17. The number of data is $D = 31$. They are Canada (CAN), France (FRA), United Kingdom (GBR), Germany (DEU), Hong Kong (HKG), Malaysia (MAS), Taiwan (ROC), Singapore (SIN), Thailand (THA), United States (USA), and Vietnam (VIE). Mainland China is not included because no data are available in some periods, and no data outside of Guangdong is reported in other periods. The total cumulative number of cases increased 10 times from $J(t_0) = \sum_i J_i(t_0) = 165$ to $J(t_{30}) = 1,846$ in these $N = 11$ regions. The fluctuation in the dataset originates in the observational noise partly, which arises from inaccurate diagnosis and the irregular delays in

¹World Health Organization, Cumulative number of reported probable cases of SARS, <http://www.who.int/csr/sars/country/en/index.html> (2003).

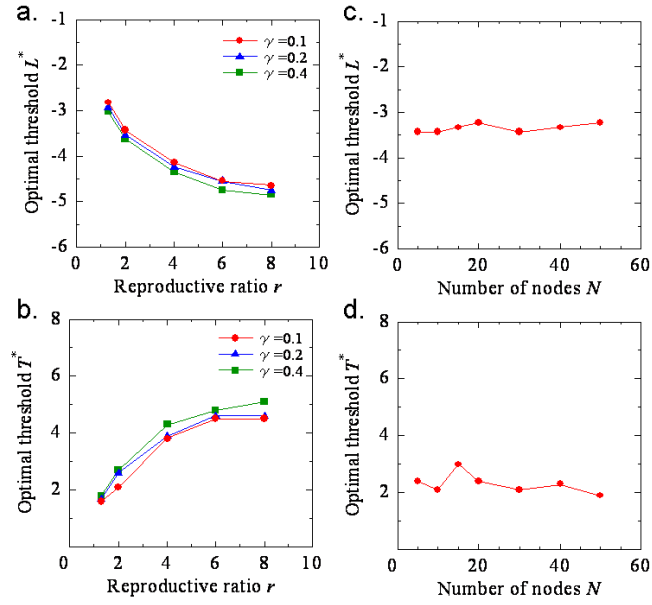


Figure 5: Optimal discrimination thresholds of the tail-end discriminator for unknown θ (a, c), and the mid-body discriminator for given θ (b, d). a, b, as a function of r and γ . c, d, as a function of N . The experimental conditions are the same as those in Figure 1 a, b.

Table 2: Test statistic L for the mid-body discriminator, T for the tail-end discriminator, and the mean m_i , variance v_{ii} , skewness s_{iii} , and kurtosis κ_{iiii} of z_i for the 11 regions where a local outbreak is reported in the early growth phase of the SARS outbreak.

| Region | L | T | m_i | v_{ii} | s_{iii} | κ_{iiii} |
|---------|-------|-----|-------|----------|-----------|-----------------|
| HKG | -41.3 | 2.1 | 0.67 | 8.7 | 0.39 | 5.1 |
| USA | -16.1 | 2.7 | -2.0 | 4.8 | 0.11 | -0.15 |
| CAN | -15.2 | 3.1 | -1.7 | 4.8 | 0.45 | 0.78 |
| SIN | -8.5 | 1.6 | -0.77 | 3.4 | -0.41 | 0.24 |
| ROC | -8.1 | 2.7 | -1.7 | 3.2 | 0.36 | -0.16 |
| MAS | -4.5 | 3.1 | -1.5 | 2.1 | 1.2 | 3.8 |
| VIE | -3.6 | 1.5 | -0.89 | 2.4 | -2.0 | 6.4 |
| GER | -2.4 | 2.1 | -0.99 | 1.4 | 0.29 | -0.81 |
| FRA | -2.4 | 2.0 | -0.85 | 1.5 | 0.70 | 0.47 |
| THI | -2.3 | 1.8 | -0.76 | 1.7 | 0.32 | 0.12 |
| GBR | -1.4 | 1.4 | -0.42 | 1.5 | 0.78 | 1.6 |
| Average | -9.6 | 2.2 | -1.0 | 3.2 | 0.20 | 1.6 |

reporting. The dataset is smoothed with a moving average filter [Walker 2010]. The window size here is $W = 3$ ($\approx 0.1D$). It is postulated that the empirical law in eq.(14) holds true.

Table 2 shows the calculated test statistic L_i for the mid-body discriminator, T_i for the tail-end discriminator, and the mean m_i , variance v_{ii} , skewness s_{iii} , and kurtosis κ_{iiii} of z_i for the 11 regions. The lower order moments averaged over the all regions are $\langle m_i \rangle = -1.0$ and $\langle v_{ii} \rangle = 3.2$. The entire dataset is disturbed. The lower order moments tend to be anomalous for United States, Canada, Singapore, and Taiwan, while the higher order moments are anomalously large for Malaysia and Vietnam. The estimated values of the parameters are $\hat{\alpha} = 0.18$, $\hat{\beta} = 0.13$ ($r = 1.4$), and $\hat{\gamma} = 0.13$. The optimal threshold $L^* = -3.6$ is obtained in the experiment for this condition. Perturbation is discovered in the time sequence data for Hong Kong, United States, Canada, Singapore, Taiwan, Malaysia, and Vietnam. The data for Hong Kong is disturbed extremely strongly. The variance and kurtosis are anomalous particularly.

Recall that Mainland China is not included in the dataset. The actual cumulative number of cases in Guangdong alone exceeded that in Hong Kong in the period. Guangdong was a spreader which had not been known to the rest of the world in the early growth phase. This is illustrated by an example of the uncovered chains of transmission. Two of the index patients in Toronto in Canada and another three of the index patients in United States stayed a hotel in Hong Kong, where a Chinese nephrologist, who had treated many patients in Guangzhou in Guangdong and become infected, was staying in late February².

²SARS Expert Committee (Hong Kong), SARS in Hong Kong: from experience to action,

This implies that China was influential potentially in infecting Hong Kong, United States, and Canada. The large amount of passenger traffic between south China and Southeast Asian countries is supposed to affect the spread in these countries. On the other hand, the disturbance is not evident in the data for European countries. It is highly probable that the discovered signal of perturbation originates in the transmission between China, a missing spreader node, and north Pacific Rim countries.

4.2.2 Flu pandemic in 2009

The flu pandemic in 2009 was a global outbreak of a new strain of the H1N1 swine influenza A virus. The virus appeared in Veracruz in southeast Mexico, in April 2009. The pandemic spread to United States and Canada immediately, and then to the South American countries, West European countries, and Pacific Rim countries. It began to decline in November. WHO archives the cumulative number of the reported laboratory-confirmed cases of the flu pandemic³. The dataset in the archive had been updated every day. It is a collection of time sequence data $J_i(t_d)$ with $\Delta t = 1$ day.

The target geographical regions in this study are those where five or more cases had been reported in about three weeks since April 28. The number of data is $D = 25$. They are Australia (AUS), Belgium (BEL), Brazil (BRA), Canada (CAN), Chile (CHL), China (CHN), Colombia (COL), Costa Rica (CRI), Ecuador (ECU), El Salvador (SLV), France (FRA), Germany (DEU), Israel (ISR), Italy (ITA), Japan (JPN), Mexico (MEX), New Zealand (NZL), Panama (PAN), Peru (PER), Spain (ESP), United Kingdom (GBR), and United States (USA). The cumulative number of cases increased 100 times from $J(t_0) = 105$ to $J(t_{24}) = 11,129$ in these $N = 22$ regions. The dataset is smoothed with a moving average filter whose window size is $W = 3$. It is postulated that the law in eq.(14) holds true for the pandemic.

Table 3 shows the calculated test statistics, the mean, variance, skewness, and kurtosis of z_i for the 22 regions. The lower order moments averaged over the all regions are $\langle m_i \rangle = -0.11$ and $\langle v_{ii} \rangle = 1.3$. The entire dataset is not disturbed. The lower order moments tend to be anomalous for United States and Mexico, while the higher order moments are anomalously large for Canada. Irregularly, Japan has large absolute values of both lower and higher order moments. The higher order moments for South American countries tend to be large while those for West European countries are small. This may be related to the seasonality for the tropics and north hemisphere. The estimated values of the parameters are $\hat{\alpha} = 0.88$, $\hat{\beta} = 0.77$ ($r = 1.2$), and $\hat{\gamma} = 0.29$. The optimal threshold $L^* = -4.2$ is obtained in the experiment for this condition. Perturbation is discovered in the time sequence data for United States, Mexico, Canada, and Japan. The ratio of this number is $4/N = 0.18$, which is remarkably smaller than $7/N = 0.64$ for the SARS outbreak. In addition, the absolute value of L

http://www.sars-expertcom.gov.hk/english/reports/reports/reports_fullrpt.html (2003).

³World Health Organization, Situation updates - Pandemic (H1N1) 2009, <http://www.who.int/csr/disease/swineflu/updates/en/index.html> (2010).

Table 3: Test statistics L , T , and the mean m_i , variance v_{ii} , skewness s_{iii} , and kurtosis κ_{iiii} of z_i for the 22 regions where a local outbreak is reported in the early growth phase of the flu pandemic.

| Region | L | T | m_i | v_{ii} | s_{iii} | κ_{iiii} |
|---------|-------|------|-------|----------|-----------|-----------------|
| USA | -17.2 | 2.4 | 0.07 | 5.2 | -0.09 | -0.55 |
| MEX | -16.9 | 1.8 | -1.7 | 4.8 | -0.31 | -0.10 |
| CAN | -8.4 | 0.84 | 0.02 | 3.4 | 2.6 | 9.6 |
| JPN | -7.3 | 3.3 | -1.5 | 2.8 | 2.3 | 5.3 |
| GBR | -2.4 | 1.1 | -0.43 | 0.74 | 0.48 | -0.56 |
| ESP | -2.2 | 1.0 | -0.05 | 0.80 | 1.4 | 2.4 |
| PAN | -1.9 | 1.1 | -0.12 | 0.92 | 1.5 | 2.2 |
| CRI | -1.6 | 1.5 | -0.15 | 1.1 | 3.0 | 9.6 |
| FRA | -1.2 | 0.93 | -0.02 | 0.67 | 1.6 | 3.6 |
| DEU | -1.0 | 1.3 | 0.21 | 0.75 | 2.0 | 3.4 |
| COL | -0.77 | 1.3 | -0.02 | 0.43 | 1.0 | 0.72 |
| ITA | -0.71 | 1.5 | 0.17 | 0.55 | 1.0 | 0.54 |
| CHN | -0.68 | 1.6 | -0.02 | 0.25 | 0.88 | 0.55 |
| CHL | -0.65 | 1.7 | 0.30 | 0.84 | 2.6 | 7.1 |
| SLV | -0.61 | 1.2 | -0.01 | 0.54 | 1.3 | 1.1 |
| NZL | -0.58 | 1.4 | 0.02 | 0.48 | 2.2 | 5.4 |
| BRA | -0.52 | 1.4 | -0.03 | 0.60 | 3.2 | 11.6 |
| ECU | -0.51 | 1.6 | 0.16 | 0.89 | 3.8 | 14.4 |
| PER | -0.35 | 1.9 | 0.15 | 0.51 | 2.3 | 6.1 |
| AUS | -0.27 | 1.7 | 0.27 | 0.64 | 1.9 | 2.4 |
| BEL | -0.24 | 1.8 | 0.14 | 0.47 | 2.1 | 3.6 |
| ISR | -0.24 | 1.8 | 0.08 | 0.34 | 0.81 | 0.04 |
| Average | -3.0 | 1.6 | -0.11 | 1.3 | 1.7 | 4.0 |

is much less anomalous than that for the SARS outbreak. In most regions, L is close to 0 while L is less than -2 for the SARS outbreak. The disturbance from an unknown origin of perturbation is small and localized. Then, what happened to the spread in United States, Mexico, Canada, and Japan?

The national transportation statistics of United States⁴ reports that most passengers cross the borders between United States and Mexico, and United States and Canada by land. The annual number of passengers from Mexico to United States by land is 24 times larger than that by air in 2009. The number from Canada to United States by land is 7 times larger than that by air in 2005. The meta-population network model relies on eq.(14) which is known valid for the world-wide air transportation. The actual heavy land transportation lets the probability of movement across the border deviate from the value estimated

⁴Research and Innovative Technology Administration, Bureau of Transportation Statistics, USA, http://www.bts.gov/publications/national_transportation_statistics (2010).

in the model. This is supported by the result that the fraction $\hat{\gamma}$ is twice as large as that for the SARS outbreak. The deviation imposes an impact on the speed of the spread at such cities near the border as San Ysidro CA, then inland cities, and finally such cities near another border as Buffalo-Niagara Falls NY [Balcan 2009]. The discovered signal of perturbation for United States, Mexico, and Canada could originate in a localized exception to the empirical law for the probability of movement, rather than a missing spreader node.

The onset of infection in Japan started suddenly on May 9 after many travelers had returned from North America during the Japan's two week long festive break. Flu spread explosively among high school students in Osaka and Hyogo. The effective reproductive ratio reached the peak value around May 14 [Nishiura 2009]. The ratio is significantly higher than 1.4 to 1.6 from an epidemiological analysis, or 1.2 from a genetic analysis, in Mexico and other countries [Fraser 2009]. School closure started on May 17 just after the secondary transmission had been confirmed and announced officially. The effective reproductive ratio declined below 1. The reproductive ratio in Japan rose and fell sharply in 10 days in early May. The precondition of the SIR compartment model is that the values of the transmission parameters do not depend on time and sub-populations. The discovered signal of perturbation for Japan could originate in an accidental localized anomaly on the probability of infection, rather than a missing spreader node.

The above results demonstrate the potential capability of the discriminator in discovering unknown origins of perturbation which violate the preconditions of the mathematical models. On the other hand, the signal discovered by the discriminators satisfies the sufficient condition to ascertain the presence of a missing spreader node incompletely, although it satisfies the necessary conditions. Prior knowledge on the demographic and socioeconomic nature of individual regions makes up for the incompleteness. *This is similar to the false positives where the signal identified an early warning for a catastrophe originates in wide classes of impending transitions because the underlying mechanism is understood incompletely [Scheffer 2009]. It is also of interest to see if the discovered location of a missing node satisfies the criterion for a strongly influential spreader in the core of the network [Kitsak 2010].* Is it possible to distinguish the presence of a missing spreader node from an anomaly on the parameters mathematically? The solution will be worked out by examining how possible origins of perturbation appear in respective order moments of the disturbed time sequence data, and by finding a suitable backward mapping from the pattern of disturbance to the possible origins. The criterion for discrimination may be a complex function of the preconditions of a mathematical model, experimental conditions, and given datasets, rather than a simple scalar threshold of a test statistic in this study. These issues are for future challenges.

5 Conclusion

This study poses an intriguing problem which few studies have addressed before. The problem is a node discovery problem for the spread of an infectious disease. Two statistical discriminators are invented to solve the problem.

The performance of the discriminators is studied with the computationally synthesized datasets. The findings are as follows. The tail-end discriminator is excellent both when the parameters is given and unknown. The mid-body discriminator is the most suitable when the parameters are given. Discovering a missing intermediate node is more difficult than a missing index node, but possible. The performance depends neither on the speed of the spread, on the size of the network, nor on the number of data much. If the cumulative number of new cases is given as a dataset, discrimination is moderately excellent, and interestingly, the tail-end discriminator works more excellently when the parameters are unknown than when they are given. The optimal thresholds for discrimination depend solely on the reproductive ratio (the probability of infection and recovery).

The WHO dataset on the SARS outbreak and flu pandemic are analyzed with the discriminators. The findings are as follows. The entire dataset on the SARS outbreak is disturbed. The signal of perturbation from an unknown origin is discovered in the time sequence data for Hong Kong, United States, Canada, Singapore, Taiwan, Malaysia, and Vietnam. The data for Hong Kong is disturbed extremely strongly. It is highly probable that the discovered signal originates in the transmission between China, a missing spreader node, and north Pacific Rim countries. In the flu pandemic, the signal of perturbation is discovered in the time sequence data for United States, Mexico, Canada, and Japan. The ratio of the number of these regions is much smaller than that for the SARS outbreak. The test statistics are much less anomalous than those for the SARS outbreak. The disturbance in the dataset is small and localized. The signal for United States, Mexico, and Canada could originate in the heavy land transportation which is an exception to the empirical law for the probability of movement. The signal for Japan could originate in an accidentally sharp rise and fall of the probability of infection.

Two advanced problems still remain unsolved. One is raising the accuracy of discrimination from the cumulative number of new cases as well as from the number of cases. The other is deriving the criteria to distinguish the presence of a missing spreader node from an anomaly on the parameters. Solving them is a milestone to the theory on the stochastic reaction-diffusion process where hidden external or internal entities are at work. These issues are for future challenges.

A Time evolution of moment

The time evolution of $I_i(t)$ is given by the Langevin equations in eq.(17). The fluctuation terms $\xi^{[\gamma]}(t)$, $\xi^{[\alpha]}(t)$, and $\xi^{[\beta]}(t)$ are Gaussian white noises. Their

explicit functional forms are unknown. The equations are applicable to any nodes including missing nodes.

$$\begin{aligned}
\frac{dI_i(t)}{dt} &= \frac{\alpha S_i(t)I_i(t)}{S_i(t) + I_i(t) + R_i(t)} - \beta I_i(t) + \sum_j \gamma_{ji} I_j(t) - \sum_j \gamma_{ij} I_i(t) \\
&+ \sqrt{\frac{\alpha S_i(t)I_i(t)}{S_i(t) + I_i(t) + R_i(t)}} \xi_i^{[\alpha]}(t) - \sqrt{\beta I_i(t)} \xi_i^{[\beta]}(t) \\
&+ \sum_j \sqrt{\gamma_{ji} I_j(t)} \xi_{ji}^{[\gamma]}(t) - \sum_j \sqrt{\gamma_{ij} I_i(t)} \xi_{ij}^{[\gamma]}(t).
\end{aligned} \tag{17}$$

In most cases, the outbreak is contained before the spread reaches equilibrium. In the early growth phase of the outbreak, $I_i \ll S_i$ and $R_i \ll S_i$ hold true. Eq.(17) becomes eq.(18) [Maeno 2010].

$$\begin{aligned}
\frac{dI_i(t)}{dt} &= \alpha I_i(t) - \beta I_i(t) + \sum_j \gamma_{ji} I_j(t) - \sum_j \gamma_{ij} I_i(t) \\
&+ \sqrt{\alpha I_i(t)} \xi_i^{[\alpha]}(t) - \sqrt{\beta I_i(t)} \xi_i^{[\beta]}(t) \\
&+ \sum_j \sqrt{\gamma_{ji} I_j(t)} \xi_{ji}^{[\gamma]}(t) - \sum_j \sqrt{\gamma_{ij} I_i(t)} \xi_{ij}^{[\gamma]}(t).
\end{aligned} \tag{18}$$

The time evolution of $J_i(t)$ is given by eq.(19).

$$\frac{dJ_i(t)}{dt} = \alpha I_i(t) + \sqrt{\alpha I_i(t)} \xi_i^{[\alpha]}(t). \tag{19}$$

Eq.(17) is equivalent to the Fokker-Planck equation in eq.(20).

$$\frac{\partial P(\mathbf{I}, t)}{\partial t} = - \sum_i \frac{\partial}{\partial I_i} \left(\sum_p a_{ip} I_p \right) P(\mathbf{I}, t) + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial I_i \partial I_j} \left(\sum_p b_{ijp} I_p \right) P(\mathbf{I}, t). \tag{20}$$

The coefficients a_{ip} and b_{ijp} in eq.(20) are given by eq.(21) and (22).

$$a_{ip} = (\alpha - \beta - \sum_{j'} \gamma_{ij'}) \delta_{ip} + \gamma_{pi}. \tag{21}$$

$$b_{ijp} = \{ (\alpha + \beta + \sum_{j'} \gamma_{ij'}) \delta_{ip} + \gamma_{pi} \} \delta_{ij} - \gamma_{ij} \delta_{ip} - \gamma_{ji} \delta_{jp}. \tag{22}$$

The moments satisfy eq.(23) through (26). The explicit functional forms of lower order moments are necessary to obtain higher order moments.

$$\frac{dm_i(t|\boldsymbol{\theta})}{dt} = \sum_p a_{ip} m_p(t|\boldsymbol{\theta}). \tag{23}$$

$$\frac{dv_{ij}(t|\boldsymbol{\theta})}{dt} = \sum_p (a_{ip}v_{pj}(t|\boldsymbol{\theta}) + a_{jp}v_{pi}(t|\boldsymbol{\theta})) + \sum_p b_{ijp}m_p(t|\boldsymbol{\theta}). \quad (24)$$

$$\begin{aligned} \frac{ds_{ijk}(t|\boldsymbol{\theta})}{dt} &= \sum_p (a_{ip}s_{pj k}(t|\boldsymbol{\theta}) + a_{jp}s_{pi k}(t|\boldsymbol{\theta}) + a_{kp}s_{pij}(t|\boldsymbol{\theta})) \\ &+ \sum_p (b_{ijp}v_{pk}(t|\boldsymbol{\theta}) + b_{ikp}v_{pj}(t|\boldsymbol{\theta}) + b_{jkp}v_{pi}(t|\boldsymbol{\theta})). \end{aligned} \quad (25)$$

$$\begin{aligned} \frac{d\kappa_{ijkl}(t|\boldsymbol{\theta})}{dt} &= \sum_p (a_{ip}\kappa_{pjkl}(t|\boldsymbol{\theta}) + a_{jp}\kappa_{pikl}(t|\boldsymbol{\theta}) + a_{kp}\kappa_{pijl}(t|\boldsymbol{\theta}) + a_{lp}\kappa_{pijk}(t|\boldsymbol{\theta})) \\ &+ \sum_p (b_{ijp}s_{pkl}(t|\boldsymbol{\theta}) + b_{ikp}s_{pjl}(t|\boldsymbol{\theta}) + b_{ilp}s_{pj k}(t|\boldsymbol{\theta}) \\ &+ b_{jkp}s_{pil}(t|\boldsymbol{\theta}) + b_{jlp}s_{pik}(t|\boldsymbol{\theta}) + b_{klp}s_{pij}(t|\boldsymbol{\theta})). \end{aligned} \quad (26)$$

The solutions of eq.(23) through (26) are given by eq.(27) through (30) when Δt is small.

$$m_i(t_{d+1}|\boldsymbol{\theta}) = I_i(t_d) + \sum_p a_{ip}I_p(t_d)\Delta t + O(\Delta t^2). \quad (27)$$

$$v_{ij}(t_{d+1}|\boldsymbol{\theta}) = \sum_p b_{ijp}I_p(t_d)\Delta t + O(\Delta t^2). \quad (28)$$

$$s_{ijk}(t_{d+1}|\boldsymbol{\theta}) = \frac{1}{2} \sum_{p,q} (b_{ijp}b_{pkq} + b_{ikp}b_{pj q} + b_{jkp}b_{piq})I_q(t_d)\Delta t^2 + O(\Delta t^3). \quad (29)$$

$$\begin{aligned} \kappa_{ijkl}(t_{d+1}|\boldsymbol{\theta}) &= \frac{1}{6} \sum_{p,q,r} \{b_{ijp}(b_{pkq}b_{qlr} + b_{plq}b_{qkr} + b_{klq}b_{qpr}) + b_{ikp}(b_{pj q}b_{qlr} + b_{plq}b_{qjr} + b_{j lq}b_{qpr}) \\ &+ b_{ilp}(b_{pj q}b_{qkr} + b_{pkq}b_{qjr} + b_{jkq}b_{qpr}) + b_{jkp}(b_{piq}b_{qlr} + b_{plq}b_{qir} + b_{ilq}b_{qpr}) \\ &+ b_{jlp}(b_{piq}b_{qkr} + b_{pkq}b_{qir} + b_{ikq}b_{qpr}) + b_{klp}(b_{piq}b_{qjr} + b_{pj q}b_{qir} + b_{ijq}b_{qpr})\} I_r(t_d)\Delta t^3 \\ &+ O(\Delta t^4). \end{aligned} \quad (30)$$

B Disturbed moment

The diagonal elements of the moments are derived for the simple network in 3.2 which consists of n_n , n_a , and n_s . Eq.(27) through (30) become eq.(31) through (34) for n_n .

$$\begin{aligned} m_n(t_{d+1}|\boldsymbol{\theta}, \gamma') &\approx I_n(t_d) + \{(\alpha - \beta)I_n(t_d) - \gamma(I_n(t_d) - I_a(t_d))\}\Delta t \\ &- \gamma'(I_n(t_d) - I_s(t_d))\Delta t. \end{aligned} \quad (31)$$

$$\begin{aligned}
v_{nn}(t_{d+1}|\boldsymbol{\theta}, \gamma') &\approx \{(\alpha + \beta)I_n(t_d) + \gamma(I_n(t_d) + I_a(t_d))\}\Delta t \\
&+ \gamma'(I_n(t_d) + I_s(t_d))\Delta t.
\end{aligned} \tag{32}$$

$$\begin{aligned}
s_{nnn}(t_{d+1}|\boldsymbol{\theta}, \gamma') &\approx \frac{3}{2}\{(\alpha + \beta)^2 I_n(t_d) + \gamma(\alpha + \beta)(2I_n(t_d) + I_a(t_d))\}\Delta t^2 \\
&+ [\frac{3}{2}\gamma'\{(\alpha + \beta)(2I_n(t_d) + I_s(t_d)) + \gamma(2I_n(t_d) + I_a(t_d) + I_s(t_d))\} + O(\gamma'^2)]\Delta t^2.
\end{aligned} \tag{33}$$

$$\begin{aligned}
\kappa_{nnnn}(t_{d+1}|\boldsymbol{\theta}, \gamma') &\approx 3\{(\alpha + \beta)^3 I_n(t_d) + \gamma(\alpha + \beta)^2(3I_n(t_d) + I_a(t_d)) + \gamma^2(\alpha + \beta)(I_n(t_d) + I_a(t_d))\}\Delta t^3 \\
&+ [\gamma'\{3(\alpha + \beta)^2(3I_n(t_d) + I_s(t_d)) + 6\gamma(\alpha + \beta)(3I_n(t_d) + I_a(t_d) + I_s(t_d)) \\
&+ \gamma^2(3I_n(t_d) + I_a(t_d) + 2I_s(t_d))\} + O(\gamma'^2)]\Delta t^3.
\end{aligned} \tag{34}$$

Eq.(27) through (30) become eq.(35) through (38) for n_a .

$$m_a(t_{d+1}|\boldsymbol{\theta}, \gamma') \approx I_a(t_d) + \{(\alpha - \beta)I_a(t_d) - \gamma(I_a(t_d) - I_n(t_d))\}\Delta t. \tag{35}$$

$$v_{aa}(t_{d+1}|\boldsymbol{\theta}, \gamma') \approx \{(\alpha + \beta)I_a(t_d) + \gamma(I_a(t_d) + I_n(t_d))\}\Delta t. \tag{36}$$

$$s_{aaa}(t_{d+1}|\boldsymbol{\theta}, \gamma') \approx \frac{3}{2}\{(\alpha + \beta)^2 I_a(t_d) + \gamma(\alpha + \beta)(2I_a(t_d) + I_n(t_d))\}\Delta t^2. \tag{37}$$

$$\begin{aligned}
\kappa_{aaaa}(t_{d+1}|\boldsymbol{\theta}, \gamma') &\approx 3\{(\alpha + \beta)^3 I_a(t_d) + \gamma(\alpha + \beta)^2(3I_a(t_d) + I_n(t_d)) + \gamma^2(\alpha + \beta)(I_a(t_d) + I_n(t_d))\}\Delta t^3 \\
&+ \gamma'\gamma^2(I_n(t_d) + I_s(t_d))\Delta t^3.
\end{aligned} \tag{38}$$

C Measured moment

The moments of z_i are measured under the same experimental conditions as those for Figure 1 **a, b, e, f**. The parameter $\boldsymbol{\theta}$ is either given, or unknown and estimated from the dataset. The following table shows the measured moments when a missing spreader node is absent. The values are the average over the all nodes and the trials for 100 different random network topologies.

| Parameter | $\langle m_i \rangle$ | $\langle v_{ii} \rangle$ | $\langle s_{iii} \rangle$ | $\langle \kappa_{iiii} \rangle$ |
|-----------|-----------------------|--------------------------|---------------------------|---------------------------------|
| Given | 0.088 | 1.0 | 0.15 | 0.22 |
| Unknown | 0.043 | 1.0 | 0.11 | 0.062 |

The following table shows the measured moments when the missing spreader node n_{10} is an index node.

| Parameter | | $\langle m_i \rangle$ | $\langle v_{ii} \rangle$ | $\langle s_{iii} \rangle$ | $\langle \kappa_{iiii} \rangle$ |
|-----------|----------------------|-----------------------|--------------------------|---------------------------|---------------------------------|
| Given | Neighboring node | 1.2 | 1.8 | 2.2 | 9.8 |
| | Non-neighboring node | -0.016 | 1.2 | 0.72 | 2.2 |
| Unknown | Neighboring node | 0.19 | 0.92 | 1.3 | 4.6 |
| | Non-neighboring node | 0.071 | 0.78 | 0.39 | 0.71 |

The following table shows the measured moments when n_{10} is an intermediate node.

| Parameter | | $\langle m_i \rangle$ | $\langle v_{ii} \rangle$ | $\langle s_{iii} \rangle$ | $\langle \kappa_{iiii} \rangle$ |
|-----------|----------------------|-----------------------|--------------------------|---------------------------|---------------------------------|
| Given | Neighboring node | 0.61 | 1.2 | 0.062 | 0.13 |
| | Non-neighboring node | -0.16 | 1.0 | 0.17 | 0.30 |
| Unknown | Neighboring node | 0.11 | 1.1 | 0.056 | 0.011 |
| | Non-neighboring node | 0.0087 | 1.0 | 0.16 | 0.10 |

References

- [Balcan 2009] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, and A. Vespignani: Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility, BMC Medicine Vol. 7, 45 (2009).
- [Baronchelli 2008] A. Baronchelli, M. Catanzaro, and R. Pastor-Satorras: Bosonic reaction-diffusion processes on scale-free networks, Physical Review E Vol. 78, 01611 (2008).
- [Barrat 2004] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani: The architecture of complex weighted networks, Proceedings of the National Academy of Sciences USA Vol. 101, pp. 3747-3752 (2004).
- [Christensen 2010] C. Christensen, I. Albert, B. Grenfell, and R. Albert: Disease dynamics in a dynamical social network, Physica A Vol. 389, pp. 2663-2674 (2010).
- [Colizza 2007] V. Colizza, and A. Vespignani: Invasion threshold in heterogeneous meta-population networks, Physical Review Letters Vol. 99, 148701 (2007).
- [Colizza 2006] V. Colizza, A. Barret, M. Barthélemy, and A. Vespignani: The role of the airline transportation network in the prediction and predictability of global epidemics, Proceedings of the National Academy of Sciences USA Vol. 103, pp. 2015-2020 (2006).
- [Dangerfield 2009] C. E. Dangerfield, J. V. Ross, and M. J. Keeling: Integrating stochasticity and network structure into an epidemic model, Journal of the Royal Society Interface Vol.6, pp. 761-774 (2009).
- [Fawcett 2006] T. Fawcett: An introduction to ROC analysis, Pattern Recognition Letters Vol. 27, pp. 861-874 (2006).
- [Fraser 2009] C. Fraser *et al.*: Pandemic potential of a strain of influenza A (H1N1): Early findings, Science Vol. 324, pp. 1557-1561 (2009).
- [Fujie 2007] R. Fujie and T. Odagaki: Effects of superspreaders in spread of epidemic, Physica A Vol. 374, pp. 843-852 (2007).

- [Hastie 2001] T. Hastie, R. Tibshirani, and J. Friedman: The elements of statistical learning: Data mining, inference, and prediction (Springer series in statistics). Springer-Verlag (2001).
- [Hufnagel 2004] L. Hufnagel, D. Brockmann, and T. Geisel: Forecast and control of epidemics in a globalized world, Proceedings of the National Academy of Sciences USA Vol. 101, pp. 15124-15129 (2004).
- [Isham 2010] V. Isham, S. Harden, and M. Nekovee: Stochastic epidemics and rumours on finite random networks, Physica A Vol. 389, pp. 561-576 (2010).
- [Kampen 2007] N. G. van Kampen: Stochastic processes in physics and chemistry. Elsevier (2007).
- [Keeling 2008] M. J. Keeling, and J. V. Ross: On methods for studying stochastic disease dynamics. Journal of Royal Society Interface Vol. 5, pp. 171-181 (2008).
- [Keeling 2004] M. J. Keeling, S. P. Brooks, and C. A. Gilligan: Using conservation of pattern to estimate spatial parameters from a single snapshot, Proceedings of the National Academy of Sciences USA Vol. 101, pp. 9155-9160 (2004).
- [Kitsak 2010] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. Makse: Identification of influential spreaders in complex networks, Nature Physics Vol.6, pp.888-893 (2010).
- [Kloeden 1992] P. E. Kloeden, and E. Platen: Numerical Solution of Stochastic Differential Equations. Springer (1992).
- [Lipsitch 2003] M. Lipsitch *et al.*: Transmission dynamics and control of severe acute respiratory syndrome, Science Vol. 300, pp. 1966-1970 (2003).
- [Lu 2010] H.-M. Lu, D. Zeng, and H. Chen: Prospective infectious disease outbreak detection using Markov switching models, IEEE Transactions on Knowledge and Data Engineering Vol. 22, pp. 565-577 (2010).
- [Maeno 2009] Y. Maeno: Node discovery problem for a social network, Connections Vol. 29, pp. 62-76 (2009).
- [Maeno 2010] Y. Maeno: Discovering network behind infectious disease outbreak, Physica A Vol. 389, pp. 4755-4768 (2010).
- [Nishiura 2009] H. Nishiura, C. Castillo-Chavez, M. Safan, and G. Chowell: Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan, Eurosurveillance Vol. 14, 19227 (2009).
- [Parshani 2010] R. Parchani, S. Carmi, and S. Havlin: Epidemic threshold for the susceptible-infectious-susceptible model on random networks, Physical Review Letters Vol. 104, 258701 (2010).

- [Potterat 1999] J. J. Potterat, R. B. Rothenberg, and S. Q. Muth: Network structural dynamics and infectious disease propagation, *International Journal of STD & AIDS* Vol. 10, pp.182-185 (1999).
- [Press 2007] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery: *Numerical recipes: The art of scientific computing*. Cambridge University Press (2007).
- [Rabbat 2008] M. G. Rabbat, M. A. T. Figueiredo, and R. D. Nowak: Network Inference from co-occurrences, *IEEE Transactions on Information Theory* Vol. 54, pp. 4053-4068 (2008).
- [Reis 2003] B. Y. Reis, M. Pagano, and K. D. Mandl: Using temporal context to improve biosurveillance, *Proceedings of the National Academy of Sciences USA* Vol. 100, pp. 1961-1965 (2003).
- [Riley 2007] S. Riley: Large-scale spatial-transmission models of infectious disease, *Science* Vol. 316, pp. 1298-1301 (2007).
- [Riley 2003] S. Riley *et al.*: Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions, *Science* Vol. 300, pp. 1961-1966 (2003).
- [Scheffer 2009] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara: Early-warning signals for critical transitions, *Nature* Vol.461, pp.53-59 (2009).
- [Simões 2008] M. Simões, M. M. T. da Gama, and A. Nunes: Stochastic fluctuations in epidemics on networks, *Journal of the Royal Society Interface* Vol. 5, pp. 555-566 (2008).
- [Small 2006] M. Small, C. K. Tse, and D. M. Walker: Super-spreaders and the rate of transmission of the SARS virus, *Physica D* Vol. 215, pp. 146-158 (2006).
- [Takeuchi 2006] J. Takeuchi, and K. Yamanishi: A unified framework for detecting outliers and change points from time series, *IEEE Transactions on Knowledge and Data Engineering* Vol. 18, pp. 482-492 (2006).
- [Taylor 1996] J. R. Taylor: *An introduction to error analysis - The study of uncertainties in physical measurements*. University Science Books (1996).
- [Walker 2010] D. W. Walker, D. Allingham, H. W. J. Lee, and M. Small: Parameter inference in small world network disease models with approximate Bayesian computational methods, *Physica A* Vol. 389, pp. 540-548 (2010).